 DOI: 10.3217/978-3-99161-050-2-204

# Decoding Text Embeddings From Functional MRI Data Using Deep Learning

**Lorenzo Tomaz[1], Judd Rosenblatt[1], Diogo Schwerz de Lucena[1*]**
*[1]Agency Enterprise, Los Angeles, CA, USA*
*1434 Abbot Kinney Blvd, Venice, CA 9029, United States. E-mail: diogo@ae.studio

*Introduction:* This study explores the relationship between brain activity (measured via fMRI from eight participants [5]) and the internal representations of language models like USE [1] and MPNet [4]. A deep neural network decoded text embeddings from fMRI voxel representations, leveraging a reading-out-loud task for alignment.

*Materials, Methods, and Results:* We analyzed the publicly available fMRI dataset [5], where nine right-handed adults read Chapter 9 of *Harry Potter and the Sorcerer's Stone* [3], one word at a time (0.5 s/word). Imaging volumes were acquired every 2 s using a 3T scanner, yielding voxel-wise time series of approximately 25,000–31,000 voxels per participant. The final data encompassed 1211 time samples per subject. Two pre-trained embedding models were considered: paraphrase-MiniLM-L6-v2 (USE) [1], which produces 384-dimensional embeddings, and all-mpnet-base-v2 (MPNet) [4], which produces 768-dimensional embeddings. Each embedding was synchronized with the fMRI time series by grouping four consecutive words into one embedding per 2 s interval, introducing a lag of eight words to account for delayed hemodynamic response [2]. Longer sequences (eight or twelve words) were also examined to assess the effect of additional context. A DNN mapped voxel intensities to text embeddings, and performance was evaluated via the cosine similarity index between predicted and actual embeddings using 5-fold cross-validation. USE and MPNet embeddings were decoded at the above-chance levels for all participants and at word-sequence lengths. USE consistently outperformed MPNet ($F(2, 70)=132.75$, $p<10^{-17}$), and longer sequence lengths improved decoding accuracy ($F(4, 70)=281.78$, $p<10^{-39}$), though there was no interaction between model type and sequence length ($F(4, 70)=0.07$, $p>0.9$).

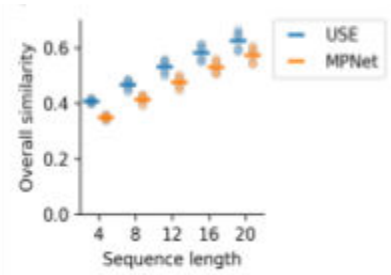


*Figure 1. Overall similarity scores across subjects as a function of the embedding model and sequence length.*

*Conclusion:* This study validates the ability to decode text embeddings from fMRI data, providing a significant step toward mapping the internal representations of large language models (LLMs) to human cortical activity. Both USE and MPNet embeddings were successfully decoded, with the former showing more substantial alignment with brain representations. These findings underscore the potential for integrating human neural data with LLMs, advancing our understanding of AI-human cognitive alignment and paving the way for future cross-disciplinary research to refine and interpret these models.

*Acknowledgments and Disclosures:* The authors express gratitude to colleagues who provided feedback. No conflicts of interest are declared.

*References*

[1] Cer, Daniel, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, et al. 2018. "Universal Sentence Encoder for English." In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 169–74. Brussels, Belgium: Association for Computational Linguistics.
[2] Holdgraf, C. R., Rieger, J. W., Micheli, C., Martin, S., Knight, R. T., & Theunissen, F. E. (2017). Encoding and decoding models in cognitive electrophysiology. Frontiers in Systems Neuroscience, 11. https://doi.org/10.3389/fnsys.2017.00061
[3] Rowling, J. K. 2012. *Harry Potter and the Sorcerer's Stone: Harry Potter Series, Book 1*. Valley View: Pottermore.
[4] Song, Kaitao, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. "MPNet: Masked and Permuted Pre-Training for Language Understanding." In *Advances in Neural Information Processing Systems*, edited by H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, 33:16857–67. Curran Associates, Inc.
[5] Wehbe, Leila, Brian Murphy, Partha Talukdar, Alona Fyshe, Aaditya Ramdas, and Tom Mitchell. 2014. "Simultaneously Uncovering the Patterns of Brain Regions Involved in Different Story Reading Subprocesses." Edited by Kevin Paterson.


 Published by Verlag der Technischen Universität Graz